# Computational Reinforcement Learning using Rewards from Human Feedback

Syed Ali Raza

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

*Doctor of Philosophy*

2018

# Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Syed Ali Raza

Date: 18/July/2018

# *Abstract*

A promising method of learning from human feedback is reward shaping, where a robot is trained via human-delivered instantaneous rewards. The existing approach, which requires numerous reward signals about the quality of agent's actions from the human trainer, is based on a number of assumptions about human capabilities. For example, it assumes that humans can provide a precisely correct feedback to an agent's action, or that they would always prefer to train an agent by means of reward signals, or that they can assess an agent's actions for any length of training.

In this thesis, we have relaxed these assumptions and have addressed two important issues which are not handled by the existing approach. First, how to compute a potential function using human feedback which can indicate the correctness of an action in terms of increasing or decreasing potential. Second, how to design training methods which cater to human preferences. Furthermore, we have identified that there are two important preferences of a human trainer in the application of reward shaping: (a) a preference to transfer knowledge by providing demonstrations and (b) a preference for short training durations. To address these issues, we have introduced three new methods of computing rewards from human-feedback.

The first method, named rewards from state preference, takes human feedback as preferences of states in terms of distance to the goal state. It removes the assumption of highly accurate evaluative feedback from the user. It computes a high-quality potential function for potential-based reward shaping from only a few human feedbacks. Using feedbacks as state preferences, a ranking model is learned which computes a complete ranking of states. These state rankings define a potential function for

potential-based reward shaping. This method learns a policy much faster than a reinforcement learner which is trained without human feedbacks.

The second method, named rewards from action labels, replaces the traditional evaluative-style feedback approach with a demonstration-style feedback approach. The method caters to the human preference of providing a demonstration. It takes human-feedback as an action label for the current state, which is similar to providing demonstrations. The agent acts using its own policy. A reward function is computed by comparing agent's action with the action label. We found that this method can be favorable to a naïve user as compared to the traditional evaluative-style feedback method.

Finally, the third method, named rewards from part-time trainers, is designed to reduce the load of a single dedicated trainer by curtailing the length of a training session. A policy is taught by a number of trainers. Each trainer provides reward signals for a small number of steps. Experiments, using online crowd, showed that the random part-time trainers can collectively train a good policy. In a survey, conducted for this method, people overwhelmingly voted in favor of the idea of training for a short duration.

Overall, this thesis contributes towards further enhancing the application scope of reward shaping. It develops three new efficient techniques of conducting reward shaping using human feedbacks of different types.

# *Acknowledgements*

First of all, I am thankful to Allah for His blessings, for the knowledge He imparted to me, and for everything.

I am extremely thankful to Prof. Mary-Anne Williams for the supervision and support she has provided throughout my candidature. I found her a courageous and lively supervisor. I am thankful to her for exposing to me a broader view of research, academia, and industry. I will be benefiting from her views in the rest of my life.

I am thankful to Prof. Sajjad Haider for his support which enabled me to begin this journey. The technical and non-technical knowledge I gained from him helped me to solve numerous problems. Specially, my technical writing style is greatly influenced by his writing style.

I am thankful to my lab members, all of them contributed something directly or indirectly to my research. Specially, I would like to thank Ben, Xun, Rony, Shaukat, Pramod, Anshar, Wei, Nima, Mahya, Sylvan, Jon, Meg, and Sam. A special thanks to Richard for proofreading this thesis and providing helpful suggestions.

I am thankful to the Australian nation and government for providing me this opportunity. It provided an exposure to world class technologies and facilities. The love for science and research here was a key motivating factor to me. Besides, I enjoyed spending time with the people, I met here, and building relationships.

I am thankful to my parents for enormous invisible support they provided. Instead of advising me to become a money making machine, they have always provided unconditional support to fulfill my desire of higher studies. My younger brother, Hasan, who helped me in supporting family in this duration. I am also thankful to my wife, Munazzah. Because of her, my productivity increased many times.

I am thankful to my friends, specially, those who instigated in me an urge to do a PhD. I am thankful to my friends in Sydney with whom I made countless fun-filled trips.

Finally, I am thankful to everyone who contributed something to this thesis.

# Contents

# List of Figures

# List of Tables

*Dedicated to my family and friends*